Inter-Speaker Variation in Spanish. An Experimental and Acoustic Preliminary Approach

Marrero, Victoria; Gil, Juana and Battaner, Elena
Universidad Nacional de Educación a Distancia
Madrid (Spain)
E-mail: vmarrero@flog.uned.es, mgil@flog.uned.es, elenabattaner@msn.com

ABSTRACT

Main factors that traditionally have been proved to influence the process of speaker identification are fundamental frequency (F0), vocalic formants, nasal energy distribution, and LTAS. The aim of this preliminary paper is to establish certain methodology in order to confirm the pertinence of these factors in Spanish. Based on a certain part of oral corpus AHUMADA, two successive procedures have been followed: first, an acoustic descriptive analysis of relevant parameters. Second, a perceptual test was designed to assess the role of these elements in the identification of speakers.

1. INTRODUCTION

An overview on the speaker-identification literature reveals three main types of research: a) those based on the visual recognition of spectrograms [1]; b) the ones concerned with the perceptual recognition of the speaker [2, 3, 4]; and c) those involved in automatic speaker identification, mainly used for the present research. A whole set of different parameters to apply to the study of voice identification is stated throughout this literature, although there are not coincident or conclusive results when approaching a hierarchy of decisive parameters. It seems necessary to find objective and specific parameters in the study of voice identification.

At least since Stevens [5], certain parameters have been employed in the study of voice identification: F0, formantic frequencies, turbulent noises, nasal consonants, etc. Parallel to the development of technical research, this set of parameters has been increased. For example, Hollien [6] added –among others-, the temporal vector, the long term average spectrum, and the vocalic formant vector (already anticipated by Ladefoged and Broadbent in 1957), which included vocalic frequencies and distances between the three first formants. This features taken as a set would provide a profile of the speaker based on *natural* features. On the relative weight of these parameters there is also an important literature, but not consensus on a priority between them. Usually, it seems basic the F0, the LTAS, or the spectral structure (formant frequencies). Nevertheless, recent research is centred in the methodological aspect of speaker identification: that is to say, in different systems (parametric or non-parametric) to manipulate and control voice quality. On the other hand, strict phonetic studies are, at this point, not so numerous.

In the selection of the parameters employed for this research, it seemed necessary to take into account certain criteria pointed out by Wolf [7]: they must be particular, natural, and frequent in normal speech, and hence, easy to measure. LTAS is usually considered as very resistant in this sense [6, 8]. It is also suggested the study of formant frequencies mean values (they can inform about the length and width of the vocalic tract) as they resist noise, time, psycho-physiological conditions. However, albeit Wolf dissuaded the study of nasals, they are commonly pointed out as very valuable [9]. The aim of this thoroughgoing research is to confirm the importance of these parameters in Spanish, devising a methodology that allow us to determine its magnitude for speaker recognition.

2. METHODOLOGY

A) DESCRIPTIVE ACOUSTIC ANALYSIS

A set of eight masculine subjects has been set up to attest former parameters' behaviour in Spanish. This set is included in Spanish oral corpus AHUMADA [10]. The reading of a phonetically balanced text task was chosen. On the theoretical bases exposed in section 1, the analysed parameters were the following:

A.1. Pitch: Mean frequency and standard deviation.

A.2. LTAS: Mean intensity (dB) and standard deviation. Spectral mean (Hz) and standard deviation.

A.3. Vowel segments:
Parameters: F0 / F1 / F2 / F3 / F0-F1 distance /F2-F3 distance / F0-F3 distance.
Variables:
- Stress: [í é á ó ú] vs. [i e a o u]
- Position within the phonic group: initial vs. final
- Context: [e a o] between front vs. back consonants

A.4. Nasals
Parameters: F1 / F2 / F3
Variables: [m] + [i a u]
[n] + [i a u]
[ñ] + [i a u]

ISBN 1-876346-48-5 © 2003 UAB

Nasal + front context ([p])
Nasal + back context ([k])

A.5. VOT

Parameters: duration
Variables: [p] + [i a o]
[t] + [i a u]
[k] + [e a u]

A.6. High resonance fricatives:

Parameters: turbulence beginning and point of maximum intensity (Hz and dB)
Variables: [s] + [é o á]
An amount of 190 measures was achieved, which means a total number of 1520 values (190 x 8).

The analysis was accomplished by Multi-Speech Signal Analysis Workstation, Model 3700, version 2.5, developed by Speech Technology Research Ltd (Kay Elemetrics Corp.). The only values obtained automatically were those related to pitch and LTAS. The remaining data were taken by the three authors of this paper from visual analysis of spectrograms.

B) PERCEPTUAL TEST

Extreme speech samples were selected in order to design the listening test. The signal was manipulated on the basis of the following criteria:

B.1. Pitch. Band reject filter, cutoff at 300 Hz; selection of samples with open vowels to avoid the appearance of F1. The sequence was added successively to build up an emission of suitable duration.

B.2. LTAS. Reversed sampled data throughout the whole emission of the text (around 50 seconds).

B.3. Vocalic space. Analysed vowels from the two subjects with higher values were trimmed without transitions. The obtained vocalic sequence was repeated for 2-3 seconds.

B.4. F3. A Band pass filter with cutoff at 2300-2500 Hz. was applied. Samples were selected from back ([o u]) in order to avoid F2 interference. Fricatives were also kept away as they could interfere the result of the perception.

B.5. Distances between formants.

a. F0-F3:
- Union of pitch and F3 filters; sequences with [a] and without high resonance.

b. F0-F1:
- Selection of the utterance with longer distance between both values (610 Hz., corresponding to [á] in "faltarme", subject 005). It was trimmed without transitions and copied to build a sequence of vowel [a] up to 2-3 seconds. This procedure was repeated with subject 014, which shows less distance in this context. Filtered above 750 Hz. in both cases.
- Selection of the performance with shorter distance between both values (250 Hz, corresponding to [ú] in "mucho", subject 008). It was trimmed without transitions and copied to build a sequence of suitable duration. The procedure was repeated with [ú] in "mucho" from subject 017, which shows longer distance in this context (383 Hz.) Filtered above 500 Hz. in both cases.

c. F2-F3:
- Selection of the utterance with higher distance in the total amount of the sample. It was compared to the shorter distance within the same context, following the procedure of former sections. Filtered above 1000 Hz. up to 3000 Hz.
- Selection of the utterance with shorter distance and comparison with the one with higher distance within the same context, following the procedure of former sections.

B.6. High resonance fricatives. Comparison of [s] with the lowest beginning and the one with the highest one. Comparison of [s] with higher maximum-intensity and lower maximum-intensity. In both cases the fricative segments were trimmed and copied to obtain a sequence of suitable duration.

Nasal consonants and VOT measures were not employed for the listening test. All these new signals were achieved and improved with Multi-Speech.

3. RESULTS

A) DESCRIPTIVE ACOUSTIC ANALYSIS
Bold = maximal value; *italic* = minimal value
Subjects = 1, 5, 8, 11, 14, 17, 20, 24.

a.1. Pitch

Pitch	1	5	8	11	14	17	20	24
Mean frq	115,79	125,99	111,04	112,2	*100,41*	120,94	**158**	134,3
std dev.	21,54	18,25	16,04	15,38	20,33	21,45	22,61	14,6

a.2. Long Term Average Spectrum

LTAS	1	5	8	11	14	17	20	24
Intens. mean[*]	5,62	3,17	1,64	7,16	1,98	4,13	*1,6*	**6,7**
std. dev	12,29	11,61	12,51	10,43	12,9	11,47	11,8	12,5
Spectral mean	269,43	260,04	*207,31*	**396,7**	243,51	308,44	295,69	278,7
std.dev	319,07	346,84	231,59	424,3	192,81	289,63	253	306

[*] LTAS intensity mean will not be used in the next steps, because any other intensity parameter has been considered; its standard deviation also multiplies the mean values.

a.3. F3

F3	1	5	8	11	14	17	20	24
mean	2485	2478	**2518**	2462	*2371*	2505	2396	2479
std dev	*101*	130	149	195	194	129	**274**	196

These measures were obtained from the mean of all vowels (18 measures for each subject). F3 mean value in the whole sample is 2462 Hz, with an standard deviation of 171 Hz. On the difference between contexts, /i/, /e/ show higher F3 values (2601 and 2532 Hz), and lower standard deviations. However, the rest of vocalic context does not follow a clear pattern: /u/ is the third vowel with higher F3, nevertheless its standard deviation is the highest. /a/, /o/ present lower F3 values (2370-2380 Hz), but considerable standard deviation, mostly in /a/ (170 Hz)

a.4. Distances between F0, F1, F2 and F3

F0-F3 distance

F0-F3	1	5	8	11	14	17	20	24
mean	2358	2350	**2397**	2342	2263	2384	*2227*	2326
std. dev.	*100*	135	154	194	197	116	**286**	202

This parameter reproduces the behaviour of the F3; in being so, it can be considered redundant.

F0-F1 distance

Mean value in close vowels ([u] [i]): 289 Hz
Mean value in medium vowels ([e][o]): 453 Hz
Mean value in open vowels ([a]): 544 Hz

Maximal value	Hz 630	= [a] subject 11⟵ minimum value in this context: 460 Hz subject 5
Minimal value	Hz 216	= [o] front context subject 20⟵ maximum value in this context = 392 Hz subject 14

Distance between fundamental frequency and first formant increases, logically, from close vowels (with a low F1) to open vowels (with a high F1). Hence it is not surprising that [a] reaches the maximal value in this chapter. Nevertheless in individual realizations the absolute minimal value does not correspond to [i] or [u], but to [o] in the front sequence [pop] -subject 20- with an extremely low F1 (400 Hz).

F2-F3 distance

Mean value in front vowels ([i][e]): 651 Hz
Mean value in medium vowel ([a]): 936 Hz
Mean value in back vowels ([o][u]): 1186 Hz

Maximal value	Hz 1700	[ú] subject 20⟵ minimum value in that context = 1267, subjects 14 and 24
Minimal value	233	[e] back context, subject 8 ⟵maximum value in that context = 634, subject 20

Distance between 2nd and 3rd formants depends on the relative position of F2, because F3 is relatively stable for each subject. In front vowels, where F2 is higher, these distance is minimal (233 Hz in [kek], subject 8); in back vowels, with low F2, it is multiplied even by 7 in our maximal value).

a.5. Openness degree and place articulation

Openness				**Front/Back**			
Subj	F1[a]	F1[i][u]	Differ	Subj	F2[i]	F2[u]	Differ
24	711	457	254	24	1933	1297	637
20	644	500	144	20	2183	1133	1050
17	632	458	174	17	1967	1084	883
14	633	408	225	14	1975	1167	809
11	693	464	229	11	1980	1145	835
8	644	433	211	8	2200	1166	1034
5	683	432	251	5	2083	1225	858
1	667	413	253	1	2114	1224	890
Mean	*664*	*446*	*218*		*2054*	*1180*	*874*

Results of degree of openness and place of articulation are inverse: subject 24, with higher vocal openness shows less variation between places of articulation; subject 20 presents more variation between front and back vowels, and he is the one with minor differences in degree of openness. First subject would show a vocalic space “long and narrow”, and the second would show a “short and wide” one.

a.6. Nasals

	F1	F2	F3
[m]	478 (70)	1322 (295)	2302 (200)
[n]	462 (52)	1364 (336)	2409 (285)
[ñ]	484 (58)	1412 (431)	2351 (429)

Mean values in Hz and standard deviation (in parentheses)

The high values of standard deviation inter and intra-speaker requires an specific analysis, so these segments are not used for the perceptual test by the moment.

a.7. VOT

Total mean: 0.024 s.
[p] mean: 0.018 [t] mean: 0.028 s. [k] mean: 0.025 s.
Max. value: 0.048 s. [ti] subject 20.
Min. value: 0.011 (various)
This elements will no be taken into consideration for the next step, due to methodological difficulties on its manipulation.

a.8. [s]

	[ése]		[ása]		[osu]		**Total**	
	begin	max.int	begin	Max.int	begin	max.int	begin	max.int
mean	1753	4337	1634	4412	1338	4092	1575	4281
std d	225	474	430	437	341	1673	372	999

The beginning of turbulence in [s] follows F2 position in the vocalic context, descending progressively from [e] to [o]. Nevertheless, the point of maximum intensity shows its higher frequency when surrounded by [a]. This parameter’s high standard deviation in back vocalic context must be emphasized.

ISBN 1-876346-48-5 © 2003 UAB

Turbulence’s beginning

Maximum Value	Hz 2416	[ása] subj. 5 ⟵min. value in this context = 1250 Hz subject 14
Minimum Value	1041	[osu] subj. 20 ⟵max. value in this context = 1583 Hz subject 5

B) PERCEPTUAL TEST DESIGN

In this section we will not offer the results obtained by listening test (that will be the aim of a forthcoming paper), but the conclusions of former paragraph applicable to the design of the test.

Subject	5	8	11	14	20	24
LTAS		m	M			
Pitch				m	M	
F3		M		m		
F0-F3		M			m	
F0-F1	m [a]		M [a]	M [o]	m [o]	
F2-F3		m[e]c.p.		m [ú]	M[e]c.p.	
Triangle					close	open
[s]	M [a] M [u]			m[a]	m[u]	

Maximum (M) and minimum (m) values

As can be seen from previous table, there are three *basic* subjects. These conform the base of the test, as three other subjects will be employed only for their results in certain parameters (the results of subjects 1 and 17 were not relevant). Taking into account time restriction and test subjects’ memory, the following formula has been accomplished:

1) Training with three voices (8, 14, and 20) not manipulated. Each voice will be identified with a photograph.

2) Assignment task of known voices. This task consists in relating the sequence with its corresponding subject (amongst the three possible). According to each parameter, the two values (higher (M) and lower (m)) will be presented.

a. Pitch: 020 (M) / 014 (m)
b. F3: 008 (M)/ 014 (m)
c. F0-F3: 020 (m) / 008 (M)
d. F2-F3 [e] in back context: 020 (M)/ 008 (m)
e. F2-F3 [ú]: 020 (M)/ 014 (m)

3) Assignment task of known (k) and unknown (u) subjects. The panel of answer will contain the three known pictures plus a black profile corresponding to the unknown voice.

a. Triangle: 020 (k) / 024 (u)
b. LTAS: 008 (k) / 011 (u)
c. F0-F1 [o]: 020 (k) / 005 (u)
d. F0-F1 [a]: 014 (k) / 011 (u)
e. [s] acute: 014 (k) / 005 (u)
[s] less acute: 020 (k) / 005 (u)

4. CONCLUSIONS

Former data are a first sketch, in broad outline, of the principal range of inter-speaker variation in Spanish with regard to the more relevant parameters mentioned in literature.

Among our findings, it is outstanding that variations in the beginning of [s] can be higher than 1000 Hz within the same context. Or that, despite Spanish vocalic system stability (especially within controlled speech –such as a reading task), differences between vocalic triangles of two subjects are higher than 400 Hz in F2 and 100 Hz in F1. Distances between formants also show an important range of variation, especially in the F2-F3 vector, albeit maintaining identical the kind of stimulus.

However, given the essentially methodological characteristics of this paper, its aims are directed towards the following step of the research: the application of the perceptual test to a set of judges. Results of this experiment will provide a tool to evaluate the relative weight of analysed parameters for speaker identification.

REFERENCES

[1] O. Tosi, *et alii* “Experiment on voice identification”, *J.A.S.A.* 51, pp. 2030-43, 1972.

[2] K. Stevens *et alii.* “Speaker identification and authentication: a comparison of spectrographic and auditory presentation of speech materials”, *J.A.S.A.* 44, pp. 1596-1607, 1968.

[3] H. Hollien *et alii*. “Perceptual identification of voice under normal, stress and disguise speaking conditions”, *Journal of* Phonetics 10, pp. 139-48, 1982.

[4] H. Kuwabara and Takagi, T. “Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method”, *Speech Communication* 10, pp. 491-95, 1991.

[5] K. Stevens. “Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds”, *Proc. 7th Intern. Congr. Phon. Sc.*, Montreal, pp. 206-27, 1971.

[6] H. Hollien. “The profile approach to speaker identification”, *Actes du XII^{ème} Congrès International des Sciences Phonétiques,* Université de Provence, Aix, pp. 396-401, 1991.

[7] J. Wolf. “Efficient acoustic parameters for speaker recognition”, *J.A.S.A.* 51, pp. 2044-56, 1972.

[8] J. Pittam. “The long-term spectral measurement of voice quality as a social and personality marker: a review”, *Language and Speech* 30, pp. 1-13, 1987.

[9] J. Glenn and Kleiner, N. “Speaker identification based on nasal phonation”, *J.A.S.A.* 43, pp. 368-72, 1968.

[10] J. Ortega, J. González, V. Marrero. “AHUMADA: A large corpus in Spanish for speaker characterization and identification”, *Speech Communication* 31, 2-3: 255-264, 2000.